Briefings in Bioinformatics, 2018, 1–10

doi: 10.1093/bib/bby056
Problem solving protocol

ATAC-pipe: general analysis of genome-wide chromatin accessibility

Zuqi Zuo, Yonghao Jin, Wen Zhang, Yichen Lu, Bin Li and Kun Qu

Corresponding authores: Bin Li, School of Life Sciences, University of Science and Technology of China, Hefei 230027, China. E-mail: libin03@ustc.edu.cn; Kun Qu, School of Life Sciences, University of Science and Technology of China, Hefei 230027, China. E-mail qukun@ustc.edu.cn

Abstract

Assay of Transposase-Accessible Chromatin by deep sequencing (ATAC-seq) has been widely used to profile the chromatin accessibility genome-wide. For the absence of an integrated scheme for deep data mining of specific biological issues, here we present ATAC-pipe, an efficient pipeline for general analysis of chromatin accessibility data obtained from ATAC-seq experiments. ATAC-pipe captures information includes not only the quality of original data and genome-wide chromatin accessibility but also signatures of significant differential peaks, transcription factor (TF) occupancy and nucleosome positions around regulatory sites. In addition, ATAC-pipe automatically converts statistic results into intuitive plots at publication quality, such as the read length distribution, heatmaps of sample clustering and cell-type-specific regulatory elements, enriched TF occupancy with motifs footprints and TF-driven regulatory networks. ATAC-pipe provides convenient workflow for researchers to study chromatin accessibility and gene regulation.
Availability https://github.com/QuKunLab/ATAC-pipe

Key words: ATAC; chromatin accessibility; transcription factor; regulatory network

Introduction

Among various DNA modulations, chromatin accessibility provides direct indications of RNA polymerases and transcription factors (TFs) occupancy, enhancer mediation and many other processes necessary for gene transcription. The altering of chromatin accessibility has been proven to be one of the key drivers of cellular functions [1–3] associating with DNA methylation, histone modifications and protein/RNA bindings. Therefore, a precise and convenient way to genome-widely survey the accessibility of different cell types under distinct conditions is required. In recent years, many techniques have been developed to detect the open chromatin sites and the positions of nucleosomes, such as FAIRE-seq [4], DNase-seq [5] and MNase-seq [6]. However, these methods often ask for large number of cells, which greatly limits their potential applications in clinical diagnosis, where way fewer number of flesh primary cells could be provided. The Assay of Transposase-Accessible Chromatin by deep sequencing (ATAC-seq) technique reported by Buenrostro *et al.* [7] used hyperactive Tn5 transposase to integrate adaptor payloads into accessible chromatin regions, which retained most advantages and the same sensitivity of endonuclease methods as DNase-seq and MNase-seq, meanwhile highly reduced the requirement of cell number and capital consumption.

Recently, numerous applications of ATAC-seq appeared on investigations of personal regulome dynamics in normal and cancer patients [8–10], cell development and differentiation [11, 12] and chromatin state kinetics [13]. Along with ATAC-seq promotion, several analysis pipelines have been proposed, such as Ataqv by Parker's lab [14], ATAC_DNas_pipelines by

Zuqi Zuo is a graduated student from Qu's Lab.
Yonghao Jing is a bachelor just graduated from Qu's Lab.
Wen Zhang is a graduated student from Qu's Lab.
Yichen Lu is an undergraduate student from Qu's Lab.
Bin Li is an associate professor at School of Life Sciences, USTC. From 2014 to 2016, he was an associate researcher in Chemistry Department of Columbia University, USA.
Kun Qu is currently a professor of Genomics and Bioinformatics at School of Life Sciences, USTC. From 2010 to 2016, he was a Bioinformatics Scientist and then the Director of Bioinformatics at Stanford University, USA.
Submitted: 20 October 2017; **Received (in revised form):** 16 April 2018

Kundaje's Lab [15] and Miskimen's standardized workflow [16], but they only cover basic processes like sequence alignment, peak calling and quality control (QC). Several toolkits adopted ATAC data for specific research interests, such as regulatory elements detection (ALTRE) [17] and TF-binding motifs prediction (DeFCoM, DeepATAC) [18, 19]. However, the great power of ATAC-seq experiments provides not only the basic information of chromatin accessibility but also the epigenomic patterns, the differential signatures of accessible regions and the effect of TF regulatory networks between samples of interest, which asks for a deeper and more systematical mining from the source data.

The absence of such a general analysis scheme for chromatin accessibility from ATAC-seq source data promoted our work on a standard program. In this article, we introduce a well-developed pipeline, ATAC-pipe, combined with several necessary bioinformatics tools and algorithms to accelerate the processing of raw ATAC-seq data for different research interests, including (1) identifying the significant differential open chromatin regions, and thereby the important sites; (2) searching for enriched binding motifs and their corresponding TFs to estimate TF effects on gene transcriptional regulation; (3) visualizing TF and motif footprints and nucleosome occupancy; (4) grouping cell-types based on correlation analysis; and (5) constructing TFs regulatory networks, to evaluate the interaction of TFs and genes among bunch of samples. Many key components of this pipeline have already been applied in our previous studies [9, 10, 13, 20], which shows its high efficiency and applicability in various research fields. The details for every procedure of ATAC-pipe are described in the next section, and the 'Results' section presents all outcomes from the raw source data to publishable figures.

Material and methods

Workflow and algorithm

ATAC-pipe requires Bowtie2, Picard, MACS2, HOMER, Python and R environment installed on LINUX/MAC OS platform, as well as the corresponding genome references and relative R package DESeq. The whole workflow presented in Figure 1 shows the priorities and relations between all procedures. For Sequence alignment, duplicate removing and peak calling, we adopted Bowtie2, Picard and MACS2, respectively; count normalization and differential analysis are carried out by DESeq and illustrated by Cluster 3.0 and Treeview; also, we quoted Homer and msCentipede for TF motifs searching and posterior probability evaluation. All other procedures, including sequence shifting, transcriptional start site (TSS) enrichment scoring, TSS distribution and TF footprint analysis, V-plot building and TFs' regulatory network comparison, are composed and performed by our computational solutions. Methods of ATAC-pipe are described by program functions as following paragraphs. More details (commands, parameters, output files, etc.) about the step-by-step procedures can be found in the 'Manual for ATAC-pipe' file in the Supplementary documents.

Raw data alignment, QC and peak calling

The pipeline starts from raw experimental result (fastq files), invoking Bowtie2 and MACS2 to map sequence to the reference genome and scan for remarkable peaks, respectively. Before alignment, the program cuts the raw sequence to the set length and trim off adapters inserted by Tn5 transposase. The trimmed fastq files are imported to Bowtie2 for sequence alignment, followed by Picard that removes all duplicate reads induced by polymerase chain reaction (PCR) in ATAC-seq library. Then ATAC-pipe concludes several parameters to evaluate the quality of each sample, including the number of raw reads, overall alignment rate, final mapped reads, final mapped rate, percentage of reads mapped to mitochondrial DNA (chrM), percentage of reads mapped to repeated regions (black list), percentage of reads filtered out by low mapping quality (MAPQ), score and percentage of PCR duplicates, which all are summarized into a QC table. Also, TSS enrichment score (reads that enriched at ± 2 kb around TSS versus the background) and read length distribution are automatically calculated and showed in high-resolution images. It is highly recommended that all low-quality samples should be excluded from the data set before next step operation.

Aligned data are stored into bed, bedGraph and bigWig format files, for the convenience to upload to the UCSC genome browser for visualization, and also the next analysis steps. In this conversion, all mapped reads are shifted $+4/-5$ bp depending on their strand, as Tn5 transposase binds DNA as a dimer and inserts 9 bp between two adaptors. Accordingly, the

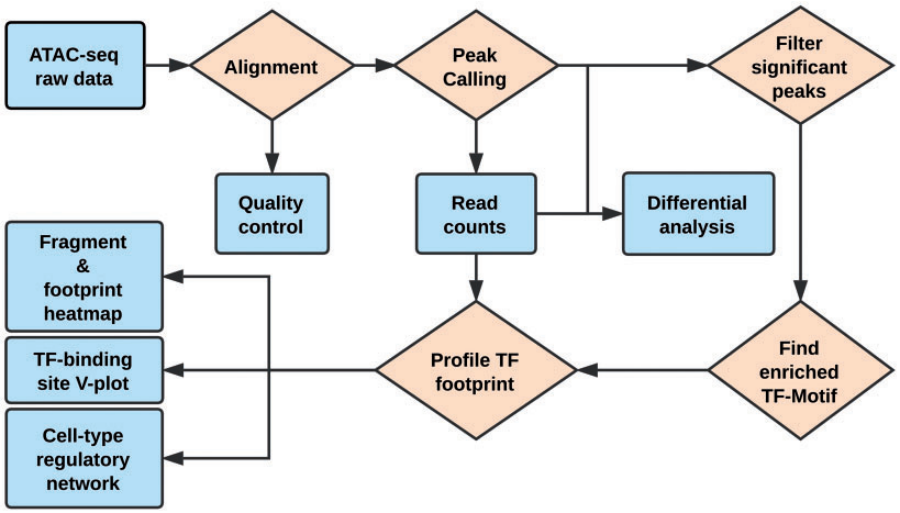

Figure 1. The scheme for ATAC-pipe: from raw data to fully demonstrated figures.

first base of each mapped read represents the Tn5 cleavage position, and all mapped reads are extended 25 bp dual sides from the cleavage positions.

Reads count and correlation analysis

ATAC-pipe merges peaks from all samples together and obtains a list of common peaks as the considerable open chromatin regions. For each sample, the numbers of raw reads residing in these peak regions are quantified. By collecting per-sample-per-peak counts for all regions, the script exports an $N \times M$ data matrix D (raw count set), where N indicates the number of merged peaks, M is the number of samples and each element $D_{i,\ j}$ means the original intensity on peak i ($i = 1$ to N) of sample j ($j = 1$ to M). Raw count set is normalized with R package DESeq, and final peak intensity is defined as logarithm of the normalized counts set. For each pair among all samples, we calculate Pearson correlation based on the log-normalized counts of all the peaks. Hierarchical clustering is carried out for the correlation matrix, and visualized by the heatmap and clustered branches yielded from Cluster 3.0.

For categories that grouped by the interactive correlations of all samples (usually analogous to their biological classification), the pipeline concatenates reads from the same group and adjusts their start sites to detect the center of the transposon's binding event (Tn5 cleavage position). Then, it counts the per base cleavage events on a genome-wide scale and produces per base signal to bedGraph and bigWig files for each category, which can then be uploaded to UCSC genome browse for tracks visualization.

Significant differential peaks

Pairwise comparison of these categories is performed by DESeq [10] to filter out significant differential open chromatin regions with discriminating parameters as *P*-value, false discovery rate (FDR), logarithm (base 2) fold change of counts and minimal peak intensity. ATAC-pipe provides scatter plot for all peaks in the frame of logarithm fold change against mean reads-per-region, colored for their differential accessibility evaluated by FDR. Meantime, it yields a filtered data matrix (significant count set, $N_{sig} \times M$, a subset of normalized count set) where N_{sig} indicates the number of significant differential peaks and M indicates the number of samples. Significant differential peaks could be binding sites of important TFs, and genes on/around these peaks are potential markers and regulatory targets with specific functions; the significant count set is applicable for hierarchical clustering with Cluster 3.0, and the heatmap of peaks' relative intensities between samples can be presented with Java Treeview.

For any category of samples, ATAC-pipe searches for known and *de novo* TF motifs with HOMER among all differential peaks in this category and ranks the most exclusive known and *de novo* motifs by *P*-values. Usually one can expect evident overlap between the known and *de novo* motifs, which indicates the reproducibility of this method. The genome-wide TF occupancy on differential accessible regions reveals the regulatory connection between the phenotype of involved sample and its corresponding TF interactions.

TSS distribution and TF footprint analysis

ATAC-pipe calculates the read length distribution with start or end site resident around all TSSs (−500 to 500 bp) and visualizes the statics data with a density profile. Per-base Tn5 transposase cleavage events around each TSSs (−500 to 500 bp) are also accumulated and showed in a heatmap permutated by the average of all nearby signals.

To analyze the footprint of each TF, the program scans TF-binding sites by invoking motif matrix with HOMER, filters out sites with specified peak list to obtain the TF-binding sites (could be a list of all merged peaks, a list of differential peaks or peak list for any cluster of interest) and counts for the per-base Tn5 cleavage events around the centers of TF-binding sites (−100 to 100 bp) [21, 22]. Meanwhile, ATAC-pipe constructs a volcano plot (V-plot) by scattering the normalized number of fragments in the picture of the fragment length versus the distance from the centers of the fragments to the center of the binding sites of the specific TF [23].

Regulatory network

From a prior score of a TF motif to its posterior probabilities among different samples (e.g. cells of patient and normal individual), ATAC-pipe adopts python version Centipede, i.e. msCentipede [24] to identify the regulatory efficiency of each motif site [25]. For each pair of TF–gene, the probabilities of the TF's motifs on the same chromosome are weighted by their relative distances of these motifs sites to the gene's TSS, and then summed up to evaluate the regulation coefficient. In this pipeline, we integrated TSS information for 28 305 genes from RefSeq GRCh37/hg19 database. A coefficient matrix with row named by TFs and column named by genes can be built for each sample, and the Pearson correlation of two samples' matrices indicates the similarity and difference between their regulatory networks. Algorithm described above was introduced by Buenrostro *et al*. [7] and proven to be reliable for the comparison of GM12878 cells and primary $CD4^+$ T cells. Here, we composed the entire procedure into ATAC-pipe and made it available for any cell types of interest. Users can easily obtain the TF-network heatmap directly from previous steps without any efforts on Centipede or other data processing.

Source data set

ATAC-seq data of $CD4^+$ T cells from normal individuals adapted in this article were downloaded from GEO: GSM2285585 and GSM2285586. ATAC-seq data of $CD4^+$ T cells from leukemia patients were downloaded from GEO: GSM2285675 and GSM2285676.

Pipeline efficiency

The ATAC-pipe has been designed as *ab initio* toolkits for the demonstration of the chromatin accessibility and downstream analysis that ends up with publishable figures. Data from different individuals can be executed parallel on multiple threads computer to accelerate the overall process. For 30 GB raw data (in fastq format) of two samples (with two replicates, respectively), using eight parallel CPU cores, it takes 2 h to map the reads and QC, 10 min to call and filter peaks and 3 h to search TF motifs and footprints. The time cost to build figures is negligible.

Comparing ATAC-pipe with existing tools

Functional comparisons between ATAC-pipe and other existing tools for ATAC-seq data analysis are listed in Table 1 and Supplementary Table S1. Most existing tools were specifically designed for one or just a few purposes of ATAC-seq data analysis, while ATAC-pipe includes more functional modules for deep data mining. For instance, ATAC_DNase_pipelines

Table 1. Comparison between ATAC-pipe and other existing tools for ATAC-seq data analysis

Tool name	Function	Input	Main output	Description
ATAC-pipe	Mapping and file processing	FASTQ group	1. Normalized BigWig file for UCSC visualization 2. Per-base or extend BED file with +4/−5 shift 3. Perbase or extend BedGraph file	Integrated pipeline with multiple toolkits for general analysis of ATAC-seq data
	QC		1. Plot of TSS enrichment score 2. Plot of fragment distribution 3. Plot and heatmap of fragment distribution around TSS 4. QC report	
	Peak calling		1. Peak with high quality selected from MACS2 output 2. Reads coverage of peaks	
	Significant regionsAnalysis		1. PCA and correlation heatmap of samples 2. File contains data significantly different between groups	
	Motif SEARCHING	Peaks	HOMER standard output with know and *de novo* found motifs	
	Footprint	BED peaks	1. Plot of motifs footprint 2. Heatmap of motifs footprint	
	Nucleosome		V-plot around given motif	
	Network	BAM TSS	TF regulation heatmap	
ATAC_DNase_pipelines	Mapping and processing	FASTQ	Normalized BigWig file for UCSC	Preliminary analysis for ATAC-seq data
	QC		1. Correlation heatmap of samples 2. QC report	
	Peak calling		1. Standard output of MACS2	
ataqv	QC	BAM TSS	1. Plot of TSS enrichment score 2. Plot of fragment distribution 3. QC report	Briefing and fast QC for ATAC-seq data
Taiji	Mapping	FASTQ group	BAM files	Integrated pipeline with ATAC-seq, RNA-seq and HiC data for TF regulatory network
	Peak calling	BED	MACS2 standard output	
	Motif Searching	Peaks	Motifs with binding site in BED format	
	Network	TFBSs[a]	Static gene regulatory network	
atactk	Footprint	BAM	Motifs footprint profile	Specific tools for motif footprint
DeepTools	Footprint[b]	BigWig BED	Plot and heatmap of coverages around given region with cluster	Visualization tools for CHIP-seq
ngs.plot	Footprint[b]	BAM, TSS	Plot and heatmap of coverages around given region	Visualization tools for CHIP-seq
NucleoATAC	Nucleosome	BAM FASTA Peak	1. V-plot around given regions 2. Plot of nuclesome occupancy around given regions	Specific tools for nucleosome occupancy
NucTools	Nucleosome	BAM BED	1. Nucleosome occupancy profile 2. Plot and heatmap of nucleosome occupancy	Specific tools for nucleosome occupancy

[a]TFBSs.

[b]For DeepTools and ngs.plot developed for CHIP-seq data, 'footprint' means signal around TSSs or given regions.

conducts raw reads mapping, sample QC and peak calling, but ATAC-pipe provides further procedures, including sample type classification, motif searching, footprint illustration, nucleosome occupancy analysis and TF regulatory network construction. More importantly, ATAC-pipe can easily identify significant regions/peaks with differential chromatin accessible sites for pairs of samples and detect enriched TF motifs in these regions, which are not applicable in other tools. In terms of their performances, with input data of 100 million reads (two samples with two replicates each, 30 GB in file size), ATAC_DNase_pipelines requires 20 h with 8-core CPU to finish mapping, QC and peak calling, while ATAC-pipe needs only 2 h to finish these steps.

For depiction of motif footprint, we compared ATAC-pipe to exiting tools atactk, DeepTools and nsg.plot. ATAC-pipe only needs one command line to produce a footprint profile, while other tools need more steps. Besides, DeepTools [26] and ngs.plot [27] were designed for ChIP-seq data, where the centers of each fragment were designated as TF-binding sites; however, a shift of +4/−5 bases from each end of the fragment was the Tn5 cleavage site, representing the chromatin accessibility. Therefore, analytical tools for ChIP-seq data may not perfectly fit to analyze ATAC-seq data. Furthermore, ATAC-pipe can automatically abstract binding sites for a motif to plot its footprint, and users only need to provide a TF name, but other tools require users to provide motif sites information beforehand.

For nucleosome occupancy analysis, both NucTools and NucleoATAC require a input file containing all targeted sequence positions defined by users [28, 29]. ATAC-pipe not only presents fragment distribution around each TSS to show nucleosome occupancy but also can inherits the motif site information for a specific TF from prior motif searching step and yield V-plot automatically illustrating the nucleus position.

For TF regulatory network, we compared ATAC-pipe with published tools Taiji, which integrates ATAC-seq and HiC data for TF regulatory network construction [30, 31]. Taiji searches transcript factor binding sites (TFBSs) with FIMO and links these sites to genes according to promoter/enhancer–gene assignments. ATAC-pipe calculates posterior probabilities for the TF's motifs on the same chromosome by Centipede algorithm and sums up TSS-distance weighted probabilities to evaluate the regulation coefficient for each TF on each gene. None procedure in our program depends on HiC data. By calculating Pearson correlation of the TF–gene coefficient matrices for multiple types of samples, ATAC-pipe illustrates the similarity and difference between regulatory effects of each TF under different cell types and automatically converted the result into a clustered heatmap.

ATAC-pipe showed better performance with higher efficiency and biologically credibility

We chose two ATAC-seq data sets from ENCODE (ENCFF019HPP and ENCFF381EMR) to compare the performance of ATAC-pipe with ATAC_DNase_pipelines (Supplementary Figure S3, Supplementary Table S2). The two tools produce almost identical results in terms of the alignment QC, TSS score and fragment length distribution (Supplementary Figure S3A–C), while ATAC-pipe is twice faster and less RAM in total but higher instant RAM, compared with ATAC_DNase_pipelines (Supplementary Figure S3D). The two tools use different strategies to filter out high-quality peaks from MACS2, which end up with different peak sets. ATAC-pipe uses *P*-value, fdr, fold change and pileup as filtering parameters, while ATAC_DNase_pipelines uses *P*-value to prescreen and then exerted pseudo-peaks as background to optimize final peaks. With default setting for both tools (*P*-value = 0.01), ATAC-pipe called out 80 000 peaks, while the other ended up with 130 000 peaks, within which 90 000 peaks overlapped with the ATAC-pipe peaks (Supplementary Figure S3E). In addition, the intensities of peaks only called out by ATAC-pipe are significantly higher than those only called out by ATAC_DNase_pipelines, suggesting the later tends to obtain more peaks with low intensity. As an example shown in Supplementary Figure S3F, peaks in green are called out only by ATAC_DNase_pipelines, and it is clear that these peaks are not distinguishable from the background.

We then chose a published ATAC-seq data set (GSE81258) of seven primary and eight metastasis tumors as a benchmarking to compare the performance of these two tools in deriving biological implications from primary cell ATAC-seq profiles. First, ATAC-pipe offers more flexible peak-calling thresholds with options of different FDR cutoffs, while ATAC_DNase_pipelines does not (Supplementary Tables S1 and S2). Second, at the same level of FDR 0.01, ATAC-pipe yielded 145 045 high-quality peaks, while ATAC_DNase_pipelines found five times more with 778 287 peaks, and all the ATAC-pipe peaks overlap with ATAC_DNase_pipelines peaks, suggesting ATAC-pipe provides more conservative peak calling than ATAC_DNase_pipelines (Supplementary Figure S4A). Third, ATAC-pipe provides a nice normalization of the peak intensities, which on average is 4-fold higher than ATAC_DNase_pipelines peaks, suggesting the latter found plenty of peaks with low intensity that were inappropriate to be considered as high-quality final peaks (Supplementary Figure S4B). We next tried to cluster the samples based on the Pearson correlations of peak intensities obtained from ATAC-pipe and ATAC_DNase_pipelines with the same FDR 0.01 cutoff. Unsupervised clustering of the samples shows that ATAC-pipe separated the 15 samples into two clearly different clusters 'hyper' and 'hypo', same as was published [32]; however, the difference between the two groups of samples was not as clear based on peaks obtained from ATAC_DNase_pipelines (Supplementary Figure S4C). Significant differential analysis between group 'hyper' and group 'hypo' also indicates a few low-intensity peaks called out by ATAC_DNase_pipelines (Supplementary Figure S4D). In the published work [32], the authors discovered *NF1* as a key TF that promotes tumor metastasis. We then used HOMER to search for the enriched motifs in the significant peaks obtained from the two pipelines and compared that with the published result. Although both pipelines enrich *NF1* motif as expected, the target TF was more enriched in ATAC-pipe peaks than ATAC_DNase_pipelines in terms of *P*-value (Supplementary Figure S4E), suggesting the former provides results with more biological credibility. In addition, we compared the time cost of ATAC-pipe and ATAC_DNase_pipelines to analyze this 15 samples with 178 Gb in file size. With 90 CPU cores, ATAC-pipe finishes the entire process in 6 h, while ATAC_DNase_pipelines takes 40 h, suggesting the former is more efficient.

Besides, we compared ATAC-pipe with atactk and DeepTools in depicting *NF1* footprint (Supplementary Figure S4F). Both ATAC-pipe and atactk showed same pattern with NFI footprint as published [32], and the former slightly outperforms in speed (25 s versus 1 min for 7821 *NF1* motif sites). DeepTools, which was originally designed to analyze ChIP-seq data and did not take a shift of +4/−5 bases from each end of the fragments in to account to precisely identify the Tn5 cleavage sites, failed to depict the footprint with a rough chart in low resolution.

Table 2. QC table of samples from leukemia patients and normal donors

Sample	Total raw reads	Overall alignment rate (%)	Final mapped reads	Final mapped (%)	chrM (%)	BlackList reads (%)	MAPQ filtered (%)	Duplicate (%)
Ctr1	181 717 542	96.81	38 445 785	21.16	30.45	0.11	12.57	32.63
Ctr2	156 966 563	96.73	27 982 417	17.83	28.82	0.08	12.89	37.2
Leuk1	196 416 559	98.10	30 278 165	15.42	51.44	0.08	11.03	20.22
Leuk2	183 067 266	98.69	19 080 768	10.42	60.59	0.05	10.49	17.19

Results

QC of ATAC-seq data

ATAC-pipe comprehensively summarizes the quality of ATAC-seq data into a report table and illustrates results with figures. According to standardized ATAC-seq experiments protocol, the biological sample should be fresh cells with high viability and no contamination of other species' genomes. The dosage of Tn5 transposase and the number of cells used also affect the quality of ATAC-seq library. ATAC-pipe takes the above aspects into consideration and provides an aggregative QC report at the end of raw data processing. In this article, we adapted released ATAC-seq data of $CD4^+$ T cells in normal and leukemia individuals as an example to present the operation of ATAC-pipe. Table 2 shows the QC results from ATAC-pipe, where we can see a high overall alignment rate, and an average 28 million final mapped reads. It suggests that the source data is unified with the genome reference and the sequencing is deep enough to call accessible regions. Meanwhile, ATAC-pipe constructed plots of read enrichment around TSS (Figure 2A) and fragment length distribution (Figure 2B). We counted TSS scores for 115 published ATAC-seq data sets (Supplementary Figure S1), announcing that high-quality data set usually has TSS score over 5.5 (with Z-score >-1). Also, the fragment length distribution for all 115 data sets was calculated by ATAC-pipe, and their envelop is the shadow region in Figure 2B. Generally, if the TSS enrichment score is <5.5, or the nucleosome position is inconspicuous or extending out of the envelop region, the data should be considered as low-quality probably caused by large ratio of dead cells and/or Tn5 insertion deficiency. For the purpose of producing cogent result through the pipeline, only high-quality data were considered on the next steps.

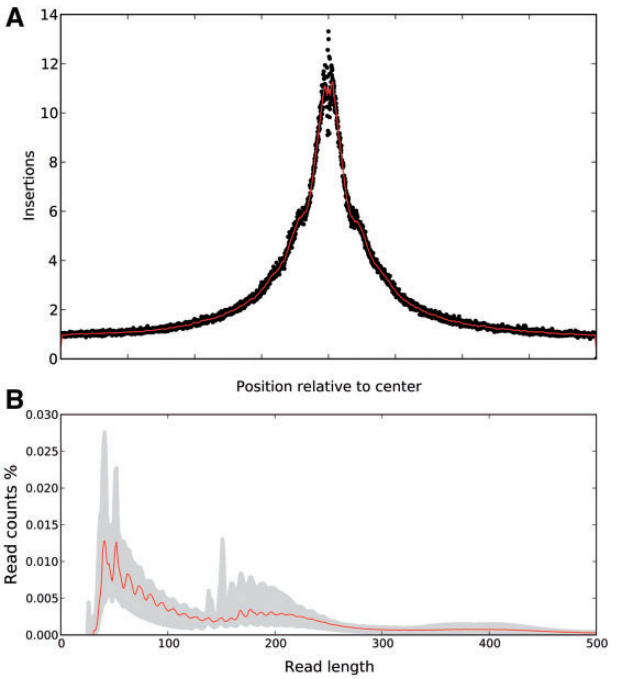

Figure 2. QC of data from ATAC-seq experiment. (**A**) Enrichment score at TSS regions (ATAC-seq data obtained from $CD4^+$ T cells from a leukemia patient). (**B**) Fragments' length distribution of the same data set (red curve) and shadows indicate the region for high-quality data.

Correlation analysis and differential peak calling

After QC screening, four high-quality samples were selected for further analysis. With Pearson correlation between normalized ATAC-seq signals of all preferred peaks for each pair of samples, ATAC-pipe yielded a hierarchical clustered heatmap to demonstrate sample groups (Figure 3A). Data of normal individuals and leukemia patients were distinctly clustered in two separated categories. Then, we compared the accessibility difference between these two categories using DESeq (Supplementary Figure S2) and calculated the logarithm fold change of normalized counts and FDR value for each accessible region (Figure 3B). To identify the chromatin signature of the normal individuals and patients, the ATAC-seq peaks are filtered with restraints, including $|\log_2(\text{fold change})|>1$, $P\text{-value}<0.01$ and $\text{FDR}<0.05$. This process ends up with 6172 differential peaks (2445 peaks enriched in patients and 3727 peaks enriched in normal donors). These peaks were then clustered using Cluster 3.0, and the relative change of their intensities was illustrated in a heatmap showing in Figure 4A.

Searching for enriched TFs in disease

Regions that are differential accessible in $CD4^+$ T cells from the leukemia patients versus those from the normal donors are potentially binding sites of key TFs that regulate the disease. We applied ATAC-pipe to search for known and *de novo* TF motifs in accessible regions/peaks that are more enriched in leukemia patients. The pipeline ranked the captured motifs for disease category by the order of *P*-values and listed the most significant ones as bar plots (Figure 4B), as is shown in the figure that the common component of the known and *de novo* motifs, i.e. *ETS*, could be a significant distinctive feature for leukemia $CD4^+$ T cells.

Nucleosome position and TF motif footprint

As DNA accessibility around gene TSS regulates gene transcription activity, it is valuable to detect the open level of TSS for all genes. Using ATAC-seq data of $CD4^+$ T cells from leukemia patients, ATAC-pipe constructed fragment distribution (Figure 5A, upper panel), which suggested nucleosome structure. Most fragments are shorter than 180 bp, which is the average length of nucleosome free fragments, and the concave between two summits shows the average distance of two nearby nucleosomes. In addition, the pipeline ranked TSS by the

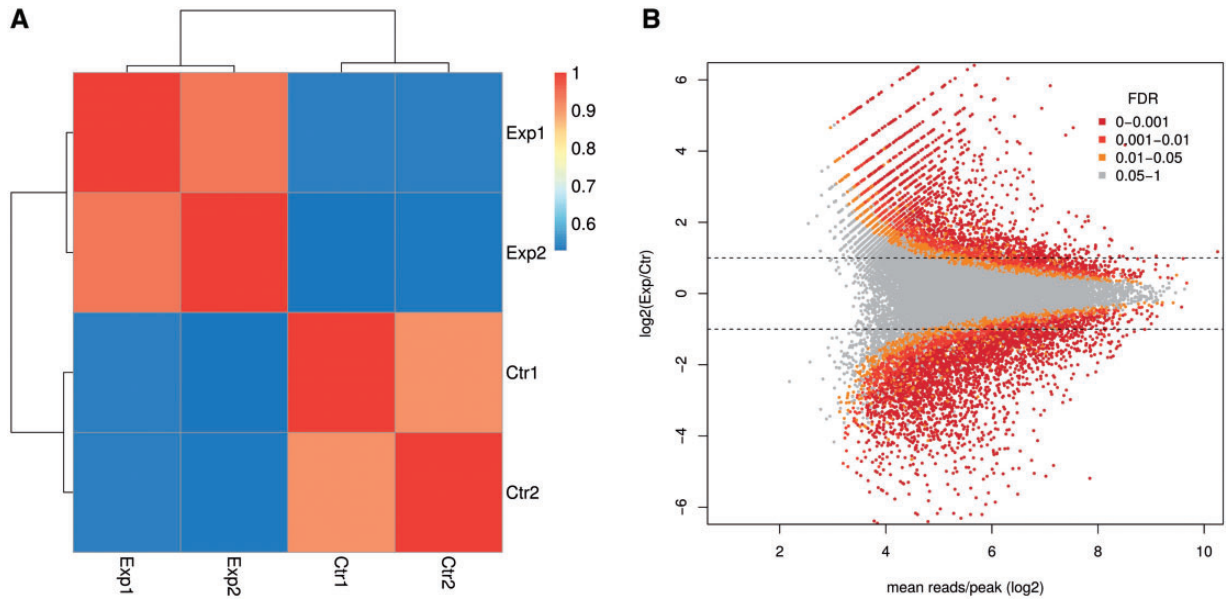

Figure 3. Differential analysis of ATAC-seq data obtained from CD4$^+$ T cells from normal individuals and leukemia patients. (**A**) Clustering of the Pearson correlation coefficients from the genome-wide ATAC-seq signals. (**B**) Differential accessibility (log2 fold change in reads per accessible region) plotted against the mean reads per region and different colors indicate levels of statistical significance measured by FDR.

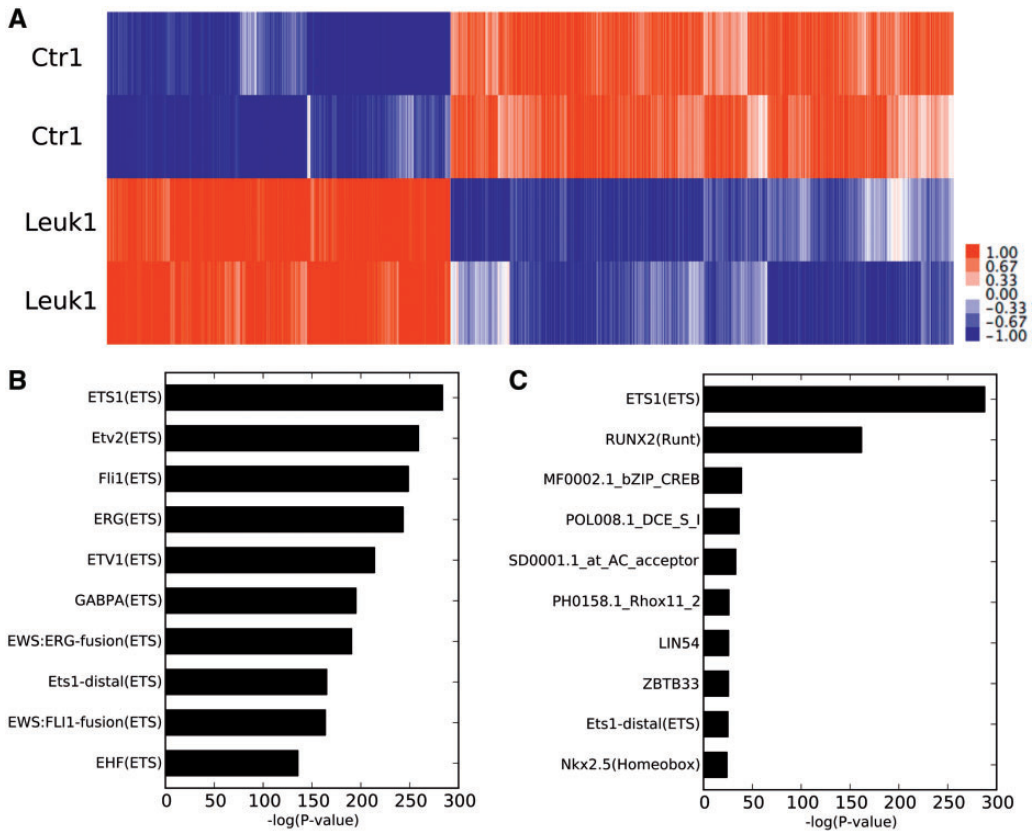

Figure 4. Comparison of chromatin states of CD4$^+$ T cells between samples from normal individuals and leukemia patients as in Figure 2. (**A**) Heatmap of the differential accessible sites. (**B**) Top-ranked known motifs that are enriched in more accessible sites in leukemia patients compared with normal controls. (**C**) Top-ranked *de novo* enriched motif.

average ATAC-seq signals around them (Figure 5A, lower panel) suggests that some genes may be transcriptional active, while others are silenced because of their corresponding chromatin states.

To illustrate the TF-binding affinity with DNA, ATAC-pipe built the footprint profile of a representative TF, CCCTC-binding factor (*CTCF*), in CD4$^+$ T cells from leukemia patients, shown in Figure 5B. After ranking the *CTCF* motif sites, ATAC-pipe built a

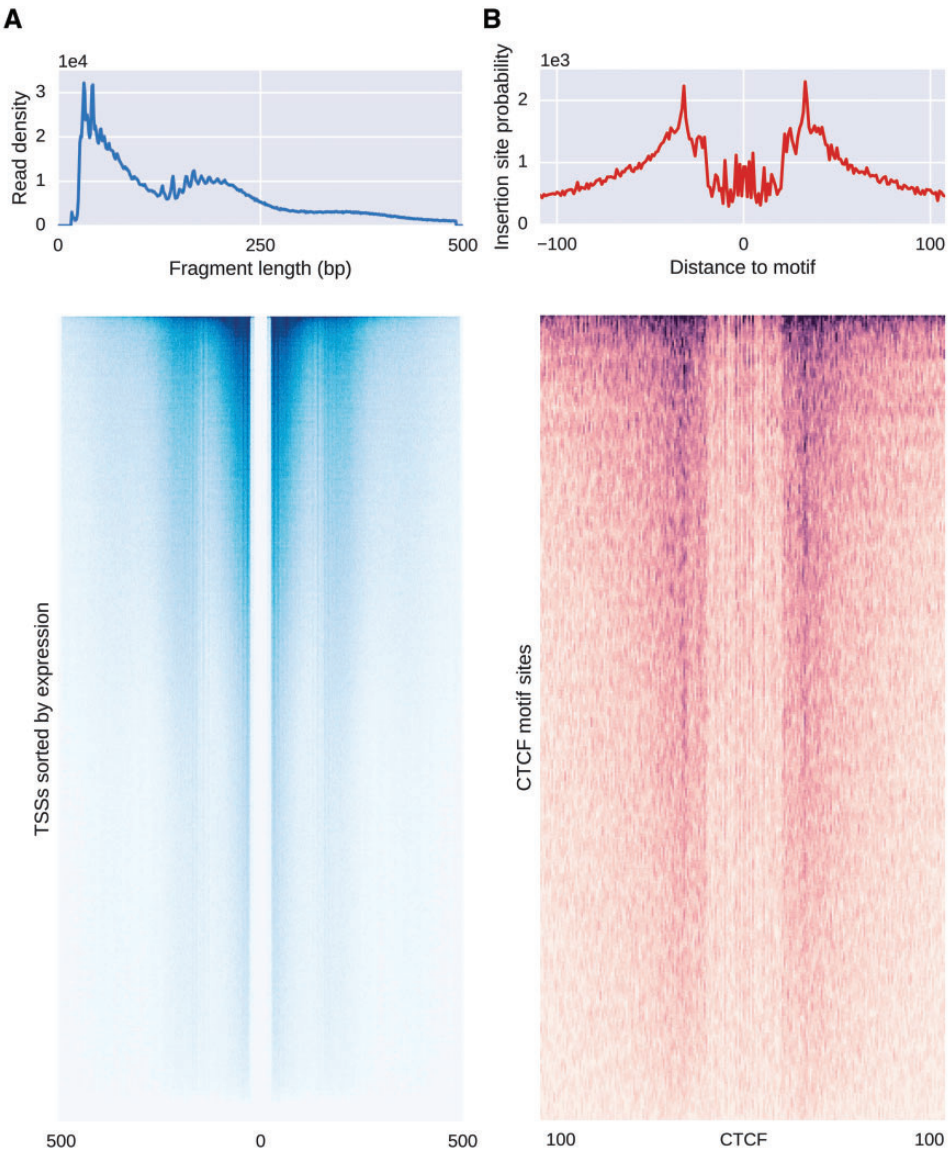

Figure 5. Nucleosome position and TF motif footprints. (**A**) Cumulative and per-site fragment length distribution for all active TSSs, ranked by average signal strength. (**B**) Aggregate footprint and per-site ATAC-seq signals at *CTCF*-binding sites.

heatmap showing ATAC-seq signal profiles and suggested diverse binding strength for different sites, indicating its heterogeneity regulatory abilities on different genes. On the other hand, to study the competition between *CTCF* and nucleosome, ATAC-pipe created a V-plot (Figure 6A) for all fragments closed to *CTCF*-binding motif sites. One dot represents the midpoint of each paired-end fragment on the plot, where Y-axis is the fragment length and X-axis is the distance from its midpoint to the center *CTCF*-binding site.

TFs regulatory network

For the TFs that refer to one's research interest, ATAC-pipe can build regulatory network by hierarchical clustering the TF–gene regulatory coefficient matrix, which is calculated from the distance-weighted posterior probabilities for each TF motif. The network demonstrated in Figure 6B captures similar (red blocks) and differential patterns (blue or weak red blocks) for $CD4^+$ T cells from leukemia patient versus normal individual. We identified several known TFs (i.e. *RUNX*, *GATA*, *STAT*, *CEBP*, *JUNAP1* and *NFKB*) that are differentially regulated in leukemia compared that with normal samples. Previous study has reported them as leukemia-specific regulatory TFs [10], which suggests that the weighed posterior probability algorithm in ATAC-pipe is proper for the estimation of TF regulatory network.

Conclusion

ATAC-pipe allows users to rapidly extract epigenomic information and construct regulatory networks from the ATAC-seq data, with highly optimized algorithms and codes. Moreover, ATAC-pipe integrated modules for statistical comparisons between multiple samples and provides publishable figures with information, including, but not limited to, significant differential accessible sites, enriched TF motifs, nucleosome positioning and dynamic regulatory networks. With the increasing interests of cellular epigenome, this toolkit can provide quick statistical analysis on regulome from dozens to hundreds of individuals

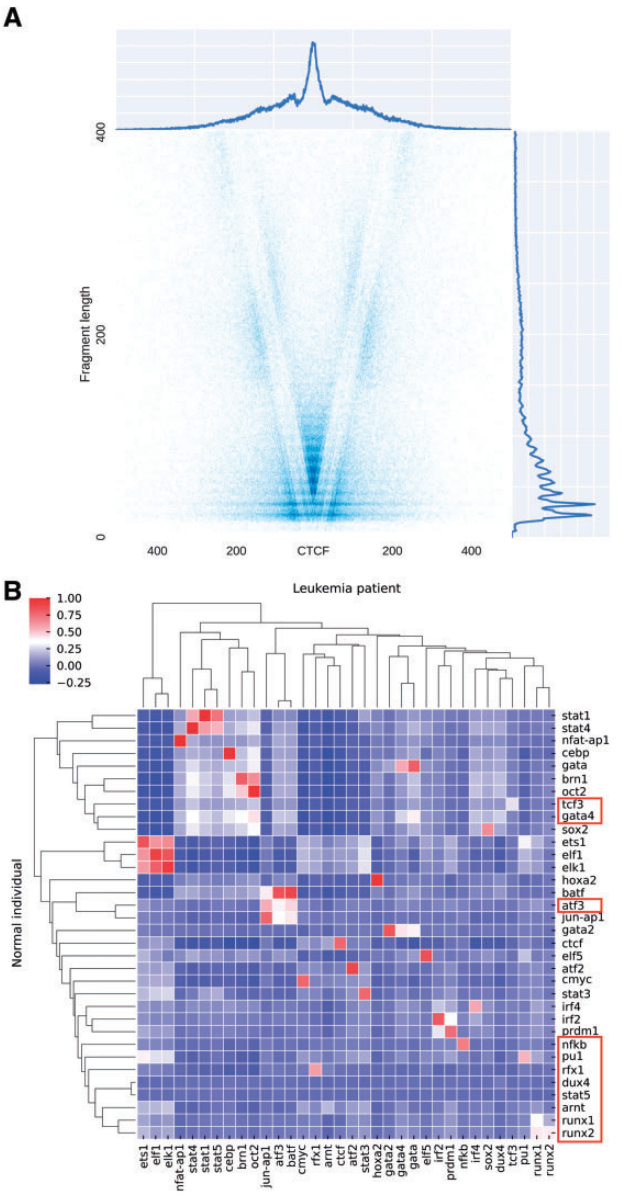

Figure 6. (**A**) V-plot for all fragments mapped around *CTCF*-binding sites. (**B**) TF-driven regulatory networks for $CD4^{+}$ T cells between leukemia patient and normal individual.

and shall dramatically reduce both experimental and theoretical expenses. Most of the key components of ATAC-pipe have already been applied in our previous studies [9, 10, 13, 20].

Results from ATAC-seq experiments have large potentials to unveil the epigenomic information and their connections with other genomic features. Combined with RNA-seq experiments with gene expression information, ATAC-pipe can provide a scheme to build complex regulatory networks. Associated with ChIP-seq data, it will also be easy to locate the specific binding sites of proteins and their repelling effects to the chromosomes. Integrated with the whole-genome bisulfite sequencing results, we can use ATAC-seq data to reveal the mechanism of the interaction between DNA methylation and chromatin accessibility. By accelerating the analysis and deep mining of ATAC-seq data, the usage of ATAC-pipe could significantly facilitate relevant biological researches in general.

Key Points

- ATAC-pipe is an efficient pipeline for general analysis of chromatin accessibility data obtained from ATAC-seq experiments and provides convenient workflow for researchers to study gene regulation.
- ATAC-pipe captures information from the quality of the original data to the genome-wide chromatin accessibility, signatures of significant differential peaks, TF occupancy and nucleosome positions around regulatory sites.
- In addition, ATAC-pipe automatically converts these results into intuitive plots at publication quality based on statistical analysis.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

The authors thank the USTC School of Life Science Bioinformatics Center for providing supercomputing resources for this project.

Funding

This work has been supported by the National Natural Science Foundation of China (grant number 81788104), the National Key R&D Program of China (grant number 2017YFA0102903; to K.Q.) and the National Natural Science Foundation of China (grant number 91640113; to K.Q.) and (grant number (31771428; to K.Q.).

Reference

1. Ho L, Crabtree GR. Chromatin remodelling during development. *Nature* 2010;**463**(7280):474–84.
2. Margueron R, Reinberg D. Chromatin structure and the inheritance of epigenetic information. *Nat Rev Genet* 2010;**11**(4):285–96.
3. Zentner GE, Henikoff S. High-resolution digital profiling of the epigenome. *Nat Rev Genet* 2014;**15**(12):814.
4. Giresi PG, Kim J, McDaniell RM, *et al*. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 2007;**17**(6):877–85.
5. Boyle AP, Davis S, Shulha HP, *et al*. High-resolution mapping and characterization of open chromatin across the genome. *Cell* 2008;**132**(2):311–22.
6. Schones DE, Cui K, Cuddapah S, *et al*. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 2008;**132**(5):887–98.
7. Buenrostro JD, Giresi PG, Zaba LC, *et al*. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 2013;**10**(12):1213–18.
8. Giorgetti L, Lajoie BR, Carter AC, *et al*. Structural organization of the inactive X chromosome in the mouse. *Nature* 2016;**535**:575–9.

9. Qu K, Zaba LC, Giresi PG, *et al*. Individuality and variation of personal regulomes in primary human T cells. *Cell Syst* 2015; **1**(1):51–61.
10. Qu K, Zaba LC, Satpathy AT, *et al*. Chromatin accessibility landscape of cutaneous T cell lymphoma and dynamic response to HDAC inhibitors. *Cancer Cell* 2017;**32**:27–41.
11. Bao X, Rubin AJ, Qu K, *et al*. A novel ATAC-seq approach reveals lineage-specific reinforcement of the open chromatin landscape via cooperation between BAF and p63. *Genome Biol* 2015;**16**(1):284.
12. Mazumdar C, Shen Y, Xavy S, *et al*. Leukemia-associated cohesin mutants dominantly enforce stem cell programs and impair human hematopoietic progenitor differentiation. *Cell Stem Cell* 2015;**17**(6):675–88.
13. Chen X, Shen Y, Draper W, *et al*. ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat Methods* 2016;**13**(12):1013–20.
14. Scott LJ, Erdos MR, Huyghe JR, *et al*. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat Commun* 2016;**7**:11764.
15. Kundaje Lab. Ataqc: Atac-seq processing pipeline. https://github.com/kundajelab/atac_dnase_pipelines.
16. Miskimen KL, Chan ER, Haines JL. Assay for transposase-accessible chromatin using sequencing (ATAC-seq) data analysis. *Curr Protoc Hum Genet* 2017;**92**:20–4.
17. Baskin E, Farouni R, Mathé EA. ALTRE: workflow for defining altered regulatory elements using chromatin accessibility data. *Bioinformatics* 2016;**33**(5):740–2.
18. Quach B, Furey TS. DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. *Bioinformatics* 2016;**33**(7):956–63.
19. Hiranuma N, Lundberg S, Lee SI. DeepATAC: a deep-learning method to predict regulatory factor binding activity from ATAC-seq signals. *bioRxiv* 2017, doi: 10.1101/172767.
20. Xu J, Carter AC, Gendrel AV, *et al*. Landscape of monoallelic DNA accessibility in mouse embryonic stem cells and neural progenitor cells. *Nat Genet* 2017;**49**(3):377–86.
21. He HH, Meyer CA, Hu SS, *et al*. Analysis of optimized DNase-seq reveals intrinsic bias in transcription factor footprint identification. *Nat Methods* 2014;**11**(1):73–8.
22. Neph S, Vierstra J, Stergachis AB, *et al*. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 2012;**489**(7414):83.
23. Kent NA, Adams S, Moorhouse A, *et al*. Chromatin particle spectrum analysis: a method for comparative chromatin structure analysis using paired-end mode next-generation DNA sequencing. *Nucleic Acids Res* 2011;**39**(5):e26.
24. Raj A, Shim H, Gilad Y, *et al*. msCentipede: modeling heterogeneity across genomic sites and replicates improves accuracy in the inference of transcription factor binding. *PLoS One* 2015;**10**(9):e0138030.
25. Pique-Regi R, Degner JF, Pai AA, *et al*. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* 2011;**21**(3):447–55.
26. Ramírez F, Dündar F, Diehl S, *et al*. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* 2014;**42**:W187–91.
27. Shen L, Shao N, Liu X, *et al*. ngs. plot: quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* 2014;**15**(1):284.
28. Schep AN, Buenrostro JD, Denny SK, *et al*. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* 2015;**25**(11):1757–70.
29. Vainshtein Y, Rippe K, Teif VB. NucTools: analysis of chromatin feature occupancy profiles from high-throughput sequencing data. *BMC Genomics* 2017;**18**(1):158.
30. Zhang K, Wang M, Zhao Y, *et al*. Systems-level identification of transcription factors critical for mouse embryonic development. *bioRxiv* 2017, doi: 10.1101/167197.
31. Yu B, Zhang K, Milner JJ, *et al*. Epigenetic landscapes reveal transcription factors that regulate CD8+ T cell differentiation. *Nat Immunol* 2017;**18**:573–82.
32. Denny SK, Yang D, Chuang CH, *et al*. Nfib promotes metastasis through a widespread increase in chromatin accessibility. *Cell* 2016;**166**(2):328–42.